Epistemic Diversity and Editor Decisions

Heesen, Remco; Romeijn, Jan-Willem

Published in:
Philosophers' Imprint

Philosophers' Imprint

VOLUME 19, NO. 39
SEPTEMBER 2019

EPISTEMIC DIVERSITY AND EDITOR DECISIONS: A STATISTICAL MATTHEW EFFECT

Remco Heesen and Jan-Willem Romeijn

University of Western Australia (RH)
University of Groningen (RH and JWR)

Abstract

This paper offers a new angle on the common idea that the process of science does not support epistemic diversity. Under minimal assumptions on the nature of journal editing, we prove that editorial procedures, even when impartial in themselves, disadvantage less prominent research programs. This purely statistical bias in article selection further skews existing differences in the success rate and hence attractiveness of research programs, and exacerbates the reputation difference between the programs. After a discussion of the modeling assumptions, the paper ends with a number of recommendations that may help promote scientific diversity through editorial decision making.

1. Introduction

The value of epistemic diversity in science has been argued extensively (e.g., Feyerabend 1975, Lakatos 1978, Longino 1990, Kitcher 1993, Hong and Page 2004, Zollman 2010). A field that harbors a greater variety of methods and theories will offer more balanced viewpoints and is better equipped to respond to challenges. In the words of Lakatos:

> The history of science has been and should be a history of competing research programmes. . . but it has not been and must not become a succession of periods of normal science: the sooner competition starts, the better for progress. (Lakatos 1978, p. 69)

In the organization of science, we should therefore aim to facilitate diversity in research programs. This holds in particular for the peer review system: A systematic bias towards a mono-culture is detrimental to scientific progress.

It is known that journal editors are prone to systematic (possibly unconscious) bias in favor of more prominent research programs (see Lee et al. 2013, pp. 9–10, and citations below). Several psychological and sociological factors underlie this tendency in editorial decision making. For instance, editors may suffer from a confirmation bias in assessing

the quality of a research program (Mahoney 1977, Ernst et al. 1992), and they may choose conservatively among the available submissions with an eye on the reputation of the journal (Resch et al. 2000, Luukkonen 2012, p. 54). But unfortunately, these are not the only drivers of bias in editorial decisions.

This paper concerns biases that are rooted not in the prejudices of editors or reviewers, but rather in the statistical characteristics of editorial decision making. Our results confront two central notions in the review process: the probability that a paper gets accepted or rejected, and the average quality of accepted or rejected papers. Comparison of different research programs with respect to these notions reveals that less well-established or otherwise vulnerable research programs are at a disproportional disadvantage. Hence, even if editors manage to purge their decision procedures of unconscious biases, they will be left with biases of a strictly statistical nature. These statistical biases contribute to the already existing tendency towards a mono-culture in science: a purely statistical Matthew effect.

Our findings on editorial decisions rely on a number of assumptions about the decision process: We presume that research papers have some latent inherent quality, that reviews offer a noisy measurement of this quality, and that editors base their decision to accept or reject a paper only on considerations of quality, informed by the reviews. In what follows we take this notion of latent quality for granted but we will return to it in our discussion.

For our first result, expounded in section 2, we further assume that there is no quality difference between the programs. However, we imagine that the editor is more familiar with the individuals, groups and networks from her own research program, and that as a consequence she has a more accurate estimation of the quality of the work. Under these minimal assumptions journal editors face a dilemma: Either they accept more papers from the research program with which they are more familiar, or the accepted papers from the more familiar program are on average of higher quality. If we add some additional assumptions, then editors fall prey to both. Assuming that editors are more likely to belong to established research programs, this makes it harder for new research programs to gain a foothold.

One possible response is that the editor should abstain from using identifying author information. Our second result, presented in section 3, shows the limits of this response. Here we assume that the programs actually differ in average latent quality, and that the more established program is also the better one. Unsurprisingly, more papers of the established program will therefore get accepted. But on top of that, the percentage of accepted papers that falls below a quality threshold is lower in the established program, no matter how this threshold is set. Moreover, the percentage of papers with sufficient quality that do not get accepted is also lower in the established program. In short, we argue that the established program enjoys more favorable error rates. This makes it once again harder for new research programs to establish themselves.

Our results identify circumstances under which a reasonable editor, who does everything in her power to choose all and only high-quality papers for publication without regard for which research program produced it, will nevertheless advantage the more established research program. Importantly, the editor treats the individual papers equally in all of this: They are judged on their quality, and on nothing else. Precisely this fairness towards individual papers leads to inequality at the level of research programs. Now, fairness towards individual papers is obviously important, but we should be aware of the non-obvious costs in terms of group-level inequalities. These mechanisms benefiting more established programs merit careful study. At a minimum, our paper aims to create awareness of them, and hence of the challenges involved in safeguarding program diversity.

One possible response is that fairness for individual papers trumps all considerations pertaining to programs, and that therefore we need not take any action whatsoever. However, we believe this response to be inadequate and aim for a different conclusion. Consider the following:

> As long as a budding research programme can be rationally reconstructed as a progressive problem shift, it should be sheltered for a while from a powerful established rival. (Lakatos 1978, p. 71)

> [W]e sometimes want to maintain cognitive diversity even in instances where it would be reasonable for all to agree that one of two theories was inferior to its rival, and we may be grateful to the stubborn minority who continue to advocate problematic ideas. (Kitcher 1990, p. 7)

We can easily multiply quotes that convey a similar preference for cognitive or epistemic diversity in science, for example from Feyerabend (1975), Longino (1990), Hong and Page (2004), Zollman (2010) and Wylie (2014). Arguably, as pointed out by Philip Kitcher (personal communication), diversity in science may not be universally beneficial, partly because dissent may have adverse effects on the role of science in public discourse (cf. Solomon 2015), and perhaps because some dissent moves beyond the confines of reasonable discussion. These caveats notwithstanding, we take the view that science benefits from diversity to be fairly widely applicable, and assume it throughout this paper. It is therefore natural to ask how we can counteract the statistical biases of peer review.

To be sure, we do not suggest to cease critical assessment of our proneness to unconscious bias, but we warn that other causes of single-mindedness are at work. If a journal is seen to promote a dominant program to the detriment of others, this cannot be ascribed simpliciter to biases at work in the editors. Instead, we should be aware that biases of a purely statistical nature may be at work in editorial decision making, and take steps to counteract them. In our conclusion (section 5), we consider what these concrete steps might be.

2. Different Familiarity with the Programs

The results of this paper rely on a basic model of peer review. We imagine a scientific community with one journal, run by an editor who decides what gets published. The members of the community produce papers which they submit to the journal. Each paper has a quality q, measured by a single real number. The editor aims to publish high-quality papers but she faces uncertainty: The quality q is unknown to the editor. When a paper arrives at the journal, all the editor has is a prior belief about its quality, in the form of a probability distribution over possible values of q.

The model thus adopts a common idea about peer review, namely that it is "the means by which one's equals assess the quality of one's scholarly work" (Eisenhart 2002, p. 241). Its aim is to guarantee "public confidence that high-quality academic work that makes a contribution to the accumulation of knowledge has been done" (Eisenhart 2002, p. 241). Conversely, bias in peer review may be defined as "any systematic effect on ratings unrelated to the true quality of the object being rated" (Blackburn and Hakel 2006, p. 378). These claims rely on a robust notion of quality, one on which it makes sense to speak of *the* quality of a journal submission. When invoked so explicitly, the notion of quality invites skepticism (cf. Lee et al. 2013). Rightfully so, we think, and we will return to this issue in section 4. Nevertheless, we take this picture of peer review to be widely, if perhaps implicitly, shared among scientists.

In her prior belief about a paper's quality, the editor takes into account the following factors. First, there are two competing research programs in the scientific community, the established research program H and the novel research program L, and each paper belongs to exactly one of these. Second, the editor is familiar with the work of some scientists in the community, but not others. The characteristics of particular scientists, insofar as the editor believes them to be relevant to the quality of their work, are represented in the model by a random variable K. If the editor has author knowledge of some kind, by knowing individual scientists, their research group, or the specific network in which they operate, then she knows these characteristics ($K = k$) and takes them into account in her prior. If the authors of a

submission are not known to the editor, she uses a generic prior that incorporates uncertainty about these characteristics.

	known author k	unknown author
research program H	$q \mid H, K = k$	$q \mid H$
research program L	$q \mid L, K = k$	$q \mid L$

Table 1: The prior distribution of q, given the research program the paper originates from and whether or not the author is known to the editor.

Submitted papers may thus be divided into four groups (known and unknown authors associated with each of the two research programs) with possibly different prior distributions (see table 1). But in the model, both of these factors are in fact irrelevant. The author characteristics follow the same distribution in the group of known scientists and in the group of unknown scientists, and the editor's beliefs are calibrated to these distributions:

$$\mathbb{E}_K[q \mid H, K] \sim q \mid H \quad \text{and} \quad \mathbb{E}_K[q \mid L, K] \sim q \mid L,$$

with $\sim$ denoting equality in distribution and $\mathbb{E}_K$ denoting expectation with respect to K. Moreover, the distribution of quality is the same for the two research programs, and the editor's beliefs reflect this as well: $q \mid H \sim q \mid L$. In sum, the editor correctly believes the papers from each of the four groups to be distributed over the quality values in the same way.

Despite all this, knowing a particular scientist's characteristics may still be relevant. For example, suppose each of the four groups consisted of just two scientists, and in each group one of these scientists consistently produces high-quality work, the other low-quality. When the editor knows the individual scientists, she can take this into account. A reasonable decision procedure might be to accept all papers from the high-quality scientist and reject all papers from the low-quality scientist. But when she does not know the individual scientists, she cannot condition her decision procedure on author identity, and she might end up making worse decisions overall. This idea drives the main result of this section.

We assume that the editor knows the characteristics of a greater proportion of scientists or research groups in the established research program H than in the novel research program L. The idea behind this assumption is that the editor has had more time to familiarize herself with the key players, the important training sites, and the essential tools and methods of the more established program. Moreover, the editor herself is typically an established member of the community, and hence she is more likely to belong to the established program. This makes it more probable that she has author knowledge for a larger proportion of papers from that program, i.e., that she is able to associate a paper with a known individual, network, or research group more easily.

The editor solicits one or more reviews of the paper. The information gleaned from the reviewers' reports is summarized in a random variable R. We assume that the quality of the paper screens off any information about the author or the research program from the reviewers' report (i.e., reviewers are unbiased with respect to these factors):

$$R \mid q \;\sim\; R \mid q, H \;\sim\; R \mid q, L \;\sim\; R \mid q, K = k.$$

The editor updates her belief about q based on the reviewers' report. Hence her posterior belief if she has author knowledge is either $q \mid R, H, K = k$ or $q \mid R, L, K = k$. If she does not have author knowledge, her posterior belief is $q \mid R, H$ or $q \mid R, L$.

Now the editor has to make a decision D whether or not to accept the paper.[1] We write $D \in \{A, \neg A\}$, where A denotes acceptance and $\neg A$ rejection. The editor aims to maximize the quality of accepted pa-

1. Since we presume that there is only one journal, strategic considerations to do with journal competition do not play a role in this decision.

pers, i.e., her utility function is given by

$$u(D) = \begin{cases} q & \text{if } D = A, \\ q^* & \text{if } D = \neg A. \end{cases}$$

This says that if the editor accepts the paper, her utility is equal to the real quality of the paper q, and if she rejects it, her utility is some fixed constant value q^*. The latter simply means that she gets no value out of rejected papers, and in particular, that she is indifferent to their quality. Very similar conclusions would be reached if we instead assumed that the editor feels regret for rejecting high-quality papers.

Since the editor does not know the quality q, she is facing a decision under uncertainty. Being a rational editor, she maximizes her expected utility. The expected utility of accepting the paper is the expected value of q, given her beliefs, i.e., it is equal to the mean of the editor's posterior distribution for the quality of the paper. The expected utility of rejecting the paper is simply q^*; no uncertainty there. So the editor accepts the paper if and only if the posterior mean quality exceeds q^*.

Given this model of editorial decisions and uncertainty, we are interested in two things. First, what is the chance that an arbitrary paper from one of the two research programs is accepted? And second, what is the average quality of published papers originating from the two research programs? We begin by discussing the results of Heesen (2018), who studies a specific instance of our model where all the relevant probability distributions are normal.

Example 1. Suppose that quality follows a normal distribution with a mean that may be different for each author and a fixed known variance: $q \mid K = k \sim N(k, \sigma_q^2)$. Suppose further that author means are themselves normally distributed in the population: $K \sim N(\mu, \sigma_k^2)$. And suppose finally that the reviewers' report provides a noisy but unbiased estimate of the quality of the paper, also with a normal distribution: $R \mid q \sim N(q, \sigma_r^2)$. If the overall acceptance rate of the journal is less than 50 % (or equivalently, if $q^* > \mu$), then the following inequalities hold (Heesen 2018, theorems 1 and 2):

$$\Pr(A \mid K) > \Pr(A) \quad \text{and} \quad \mathbb{E}[q \mid A, K] > \mathbb{E}[q \mid A].$$

That is, papers written by authors known to the editor are (on average) more likely to be accepted than papers written by unknown authors, and (despite this) the average quality of papers written by known authors that are accepted for publication is higher than the average quality of accepted papers written by unknown authors. If, as we have assumed, the editor knows a greater proportion of scientists from research program H than from research program L, it follows that the same inequalities hold at the level of research programs, i.e.,

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \text{and} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

The results from this example are worrying. They show that an editor who only aims to maximize the quality of accepted papers may accept papers from the established research program at a higher rate than those from the novel research program. Moreover, the papers she accepts from program H are of higher quality (on average) than the papers she accepts from program L.

Notice the minimal assumptions under which this result holds: Apart from the different levels of information she has about authors, the editor treats each paper equally. Various ways of making the model more realistic are likely to exacerbate the result, e.g., if the editor is biased in favor of the research program she is more familiar with. Moreover, we have assumed that the extra information the editor has only affects her assessment of the quality of individual papers. If she also finds it easier to identify good reviewers for papers she has more information about (i.e., σ_r^2 is lower when assessing papers by known authors), this would likewise exacerbate the result.

Heesen (2018) goes on to discuss whether this phenomenon produces epistemic injustices for individual authors, and the extent to which triple-anonymous peer review may avoid such injustices. He concludes that triple-anonymization (where the editor does not know the identity of the author) is advisable from the perspective of fairness, but may not be desirable from an epistemic perspective. Here, our focus is on the effect on entire research programs rather than individual authors. As we will see in section 3, we are also somewhat skeptical of the epistemic benefits of triple-anonymization, if for different reasons.

As we pointed out, however, Heesen's results depend on assuming that various uncertainties follow normal distributions. Our theorem, a partial generalization of Heesen (2018, theorems 1 and 2), shows that these results are not merely a peculiarity of normal distributions. It says that regardless of the distributions of q, K, and R, at least one of Heesen's inequalities must hold.

Theorem 2. *If knowing author characteristics sometimes makes a difference to the editor's decision (i.e., there is a positive probability of getting a combination of author characteristics and reviewers' report such that the paper is accepted if the editor has author knowledge, but rejected if the editor does not have author knowledge, or vice versa), then*

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \textit{or} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

The proof is given in appendix A. It is based on the value of information theorem due to Good (1967). The idea is that the additional information the editor has available when she has author knowledge allows her to make better decisions. (While we have framed things in terms of an established program and a novel program to highlight our concerns about epistemic diversity, the mathematical result is indifferent to this: In any situation of asymmetric information — including a situation where an editor knows more about a novel research program — the decision-making process studied here would favor the side about which more information is available.)

The theorem shows that at least one of the following holds. Either the acceptance rate for papers from research program H is higher than the acceptance rate for papers from research program L, or accepted papers from program H are on average of higher quality than accepted papers from program L. This is so even though the overall distribution of quality is the same in the two programs. We may formulate the result as a dilemma that the editor faces: Either she will be seen to display a kind of favoritism by accepting papers from the established research program at a higher rate, or she will find that the papers she publishes from the established program turn out to be better papers (on average) than those she publishes from the novel program. In other words, the dilemma is between boosting research program H directly by giving it more exposure, or indirectly by creating the misleading impression that it produces higher quality work. By adapting her editorial practices, she might manage to avoid one of these problems, but she cannot avoid both.

This is what we call a statistical Matthew effect: The established research program receives a boost despite its quality distribution being identical to that of the novel research program, and despite the fact that neither the editor nor the reviewers are biased. It is a Matthew effect (in the sense of Merton 1968) because the research program already enjoying a good reputation receives greater benefits when it delivers the same quality of work. It is statistical because it arises from the underlying uncertainties in measuring quality as opposed to a specific preference from the editor or the reviewers.

3. Latent Differences between the Programs

A salient feature of the model presented in the previous section is that the editor treats papers differently depending on their author. By and large, scientists with a good track record will have their papers accepted even if the reviewers' report is relatively lukewarm, whereas scientists with a poor track record need a glowing report for acceptance. In response to this, we may want to rule out the use of prior

information by the editor. This could be achieved by implementing triple-anonymous review.[2]

Triple-anonymous review comes at a cost: We give up information that is potentially relevant for evaluating paper quality. This is true in our model — the editor does best in selecting for quality if she factors in whether she knows the author — and it also seems to be confirmed empirically by Laband and Piette (1994).

One might advocate triple-anonymous review to prevent various other types of biases (see Heesen 2018, Lee and Schunn 2010, p. 7). However, it is not entirely successful as a response to the statistical Matthew effect. Similar phenomena can still occur if the quality distributions of the two research programs are different.

To show this, we present an adapted version of our model. As before, each paper has latent quality q and papers belong to one of two research programs (H and L). The reviewers' report R provides information about the quality of the paper, and it does so in a way that is independent of the research program that the paper belongs to: $R \mid q, H \sim R \mid q, L$. But we no longer distinguish between known and unknown authors or other such prior information: The decision to accept a paper for publication is based exclusively on the reviewers' report. In particular, the paper is accepted if R exceeds a threshold value r^*.

The reviewers' report R is a random variable which follows some probability distribution. We make no assumption on the shape of this distribution. We only assume that papers of higher quality have a greater chance of being accepted, in the following sense. Define the acceptance function a as the chance of acceptance given the latent quality q, i.e.,

$$a(q) := \Pr(A \mid q) = \Pr(R > r^* \mid q).$$

We assume that this function is (strictly) increasing, i.e., $q < q'$ implies $a(q) < a(q')$.

While we make essentially no assumptions on the distribution of R, we do make some more substantial assumptions on the distribution of the latent quality. Let F_H be the distribution of quality for papers out of research program H, that is, F_H is the function such that $F_H(x) = \Pr(q \leq x \mid H)$. Similarly, let F_L be the quality distribution for research program L. We assume that the quality distributions are differentiable so that the density functions f_H and f_L are well-defined everywhere.

We make two key assumptions on the quality distributions: One on what they have in common, and one on how they differ. First, we restrict our attention to distributions whose density function is log-concave. A density function f is log-concave just in case

$$f(px + (1-p)y) \geq f(x)^p f(y)^{1-p}$$

for all $x, y \in \mathbb{R}$, and for all $p \in [0,1]$. Log-concavity is a somewhat technical assumption restricting the shape of the distribution; among other things, it entails that the distribution is unimodal. It is satisfied by a wide range of well-known distributions, such as the normal, exponential, and uniform distributions.

Second, we assume that the quality distributions for the two research programs have the same functional form, but that the average quality of papers produced by research program H is higher than the average of research program L. The idea is that the established program is able to reliably produce work of decent quality, whereas the novel program may suffer from startup problems. This assumption need not always be satisfied, but here we explore cases where it holds, much like previously we assumed that the editor is more likely to have author knowledge for papers coming from the established program.

2. We deliberately avoid the terminology of "blind review", which has been criticized for being ableist (Tremain 2017, pp. 32–33).

More formally, let f be a log-concave density function supported on an interval $[b, c]$,[3] and let F be the corresponding distribution function. Our assumption requires that there exist parameters μ_H and μ_L with $\mu_H > \mu_L$ such that

$$F_H(q) = F(q - \mu_H) \quad \text{and} \quad F_L(q) = F(q - \mu_L).$$

So we require that quality follows the same log-concave distribution in both research programs, differing only in that the distribution for research program H is shifted to the right compared to the distribution for research program L.

We discuss the role of these assumptions in more detail at the end of this section, and we provide a more extensive critical discussion of the model in section 4. For now we note that, analogous to section 2, the assumptions of log-concavity and different average quality may be interpreted either as genuine features of the distribution of quality in the scientific community, or as features of the editor's beliefs about how quality is distributed.[4]

Our main result for this version of the model relates the probability that a paper is accepted for publication to the probability that its latent quality exceeds some threshold t. For interpreting the result, it is useful to think of the condition $q > t$ as asserting that the paper passes some threshold of suitability for publication. We may then think of the goal of selecting for quality in terms of error rates: A false positive occurs when an unsuitable paper ($q \leq t$) is accepted for publication, and a false negative occurs when a suitable paper is rejected. The theorem says that regardless of the choice of threshold t, both error rates are lower for research program H. Or, in terms of concepts from the literature on psychological testing, both the sensitivity $\Pr(A \mid q > t)$ and the positive predictive value $\Pr(q > t \mid A)$ of editorial decisions are better for program H.

Theorem 3. *Let $t \in \mathbb{R}$ be any number in the support of f_H or f_L, i.e., $b + \mu_L < t < c + \mu_H$. Then*

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L) \quad \textit{and}$$
$$\Pr(A \mid q > t, H) \geq \Pr(A \mid q > t, L).$$

The latter inequality is strict unless the right tail of F is exponential, i.e., $f(q) \propto \exp\{-q\}$ for all $q > t + \mu_L$.

A proof of the theorem is given in appendix B. It generalizes results obtained in a different (psychometric) context by Borsboom et al. (2008).

The intuition behind the proof is as follows. For any fixed quality q, the chance of acceptance does not depend on the research program, and the higher q is, the higher the chance of acceptance will be. As a result, papers that are close to the suitability threshold t are at greatest risk of an error: Those just above the threshold are less likely to be accepted than those far above it, and those just below the threshold are more likely to be accepted than those far below it. The distributional assumptions entail that among the suitable papers from research program L, there are proportionally more papers close to the threshold than among suitable papers from research program H.

Consider what this means for the novel research program. Of course, given its lower average quality, its overall acceptance rate is lower (see corollary 4 below). This is presumably as it should be. But the higher rate of false negatives means that when the novel research program produces a good paper ($q > t$), it is relatively more likely to be rejected by the journal. And conversely, the higher rate of false positives means that when the novel research program manages to get a paper accepted for publication, it is relatively more likely to be of low quality ($q \leq t$). Researchers forming an opinion of the novel program

3. This means that $f(x) = 0$ if $x < b$ or $x > c$. We explicitly allow for the possibility that $b = -\infty$ and/or $c = \infty$.

4. While features of the quality distribution may reflect the editor's perception, the quality of individual papers cannot be straightforwardly interpreted as mere editor perception, as we have assumed throughout that the editor faces uncertainty about quality. We return to the interpretation of individual paper quality in section 4.

will quickly lose faith, pointing out that despite the editor's exclusive focus on latent quality, papers from the novel program have a harder time in the review process and are more often disappointing in content when they do make it through.

The peer review we have modeled is "unbiased" in the following sense: Papers of equal quality have the same chance of being accepted regardless of the research program they originated from. Theorem 3 shows that such a peer review system may still be "biased" in the following sense: Papers whose quality exceeds a threshold value may have different chances of acceptance depending on the research program they originated from. We can recognize a continuous version of Simpson's paradox: For every subset of papers with a given quality q, there is no dependence of acceptance on the program, but owing to a different distribution over quality for the two programs, acceptance does seem to depend on program once we coarse-grain towards the binary variable of suitability, $q > t$ and $q \leq t$.

Notice that the situation is fully symmetrical and that we can therefore also derive that $\Pr(q < t \mid \neg A, L) > \Pr(q < t \mid \neg A, H)$, i.e., the negative predictive value is better for L than for H: Among the rejected papers from the more established program, there are more papers that are in fact suitable than among the rejects of the novel program. Similarly, we can derive a better specificity for L, namely $\Pr(\neg A \mid q < t, L) \geq \Pr(\neg A \mid q < t, H)$, meaning that the percentage of accepted papers among the unsuitable ones is higher in the more established program than in the novel one. The general conclusion we might therefore draw is that the programs have different error rates for acceptance and rejection, and hence that they are not treated on a par.

However, we believe we can make our conclusions more specific. The latter two errors, which are larger for program H than for program L, do not harm the reputation or the attractiveness of program H in the way that the errors in theorem 3 harm L. For one, the papers that are not accepted simply do not see the light of day. The fact that in the set of rejected papers from H a higher percentage will have the requisite quality for publication will not deter talent or make the program H look degenerate. If there were a possibility to check those papers out, the impression might become that we only see part of all the high-quality work from H. Moreover, the fact that in the pool of unsuitable papers from H, more will make it to publication by sheer luck is not damaging to program H either, as the quality of accepted papers from H is still better on the whole.

For expository purposes, we have explained theorem 3 in terms of a notion of suitability for publication. But it bears repeating that the theorem holds regardless of the choice of the threshold value t. So it follows from the theorem that the probability distribution of quality for those papers from research program H that get accepted for publication stochastically dominates the quality distribution for accepted papers from research program L: For any t, the probability that the quality of an accepted paper from program H is at least t is greater than the probability that the quality of an accepted paper from program L is at least t.

Accordingly, researchers who see only what gets published will find that the novel research program consistently underperforms. Even among papers that appear in print, papers from the novel research program are consistently worse in expectation than papers from the established research program. As a result, researchers may even (falsely) suspect the editor of applying positive discrimination in favor of the novel research program: How else to explain the consistent difference in quality even among papers deemed publishable by the editor? Thus, we claim, there is a sense in which the peer review system seems biased against the novel research program even when we take into account the fact that its average quality is lower. This arises not from any individual bias at the level of the editor or the reviewers, but from the underlying probability distributions. This is the sense in which a statistical Matthew effect operates in this second version of our model.

A few final remarks on this model. First, it may be helpful to phrase our result in a way that makes for a straightforward comparison with the results of the first model. Recall that, by theorem 2, either the acceptance rate or the average quality of published papers is higher for the

established research program. Theorem 3 entails that both inequalities are satisfied in the present version of the model.

Corollary 4. *In the model of this section,*

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \textit{and} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

Moreover, the first inequality holds for any density function f, i.e., does not require the assumption of log-concavity.

Second, we may ask what happens when the distribution of quality differs between the two research programs in both mean and variance. We may again generalize (and slightly correct) results from Borsboom et al. (2008) to obtain a partial answer for the case where research program H has the greater variance.

Theorem 5. *Define f and F as above. Let*

$$F_H(q) = F\left(\frac{q-\mu_H}{\sigma_H}\right) \quad \textit{and} \quad F_L(q) = F\left(\frac{q-\mu_L}{\sigma_L}\right).$$

Let $t \in \mathbb{R}$ be any number such that $\min\{\sigma_L b + \mu_L, \sigma_H b + \mu_H\} < t < \max\{\sigma_L c + \mu_L, \sigma_H c + \mu_H\}$. If

$$\sigma_H > \sigma_L \quad \textit{and} \quad \frac{\mu_H - t}{\sigma_H} \geq \frac{\mu_L - t}{\sigma_L}$$

then

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L) \quad \textit{and}$$
$$\Pr(A \mid q > t, H) > \Pr(A \mid q > t, L).$$

See appendix B for a proof. A similar proof can be given about the negative predictive value and the specificity of the editorial decisions, which again point in the opposite direction assuming that program L has the greater variance (cf. Borsboom et al. 2008, appendix C).

Third, we note that our model in this section differs from that of the previous section, requiring log-concavity and a difference in the means of the quality distributions for the two research programs. How restrictive are these assumptions? Their role in the proof is to guarantee a certain smoothness of the distributions, so that the result works out the same way for all values of t. The proof suggests, however, that acceptance and suitability will typically come apart for different distributions of quality. We conjecture that as long as the distributions of quality in the two research programs are different (in any which way), it is unlikely that the probabilities of suitability given acceptance will be equal, and likewise for the probabilities of acceptance given suitability.[5] At a minimum though, theorem 3 establishes that in a non-trivial range of circumstances, triple-anonymous review is vulnerable to a statistical Matthew effect, and hence that anonymity cannot be taken as a panacea against the problems raised in the previous section.

4. Discussion of Modeling Assumptions

The upshot of the foregoing is that editorial decision making is liable to purely statistical biases, and that these biases work against the diversity of research programs within a discipline. Since we take epistemic diversity to be beneficial, these biases are detrimental and we therefore need to counteract them. But can we claim that our models are sufficiently similar to editorial practice, so that we are warranted in believing that these biases indeed occur, and justified to take action against them? In what follows, we critically assess our models and answer the above questions affirmatively albeit tentatively, because ultimately, empirical study has to settle the matter.

Both models posit a latent quality of papers that is then measured by the editor. It is not immediately clear that measuring paper quality is what editors and reviewers are doing. Editorial practice consists in

5. More specifically, we conjecture that this is a measure zero event, i.e., for any F_H and for any non-trivial choice of t, the set of F_L such that $\Pr(q > t \mid A, H) = \Pr(q > t \mid A, L)$ or $\Pr(A \mid q > t, H) = \Pr(A \mid q > t, L)$ will be measure zero in the set of all probability distributions.

accepting and rejecting papers, and referee reports employ grades, often accompanied by a narrative. However, we believe that the practice of the peer review process — assessing papers for their "suitability" for publication — implicitly commits editors and reviewers to some version of our story. Our idea is that the notion of a latent, unidimensional paper quality is effectively induced by the editorial practice, or at least that such a notion will prove useful in a representation of that practice.[6] But this is ultimately an empirical matter and not one we can settle in this paper. For present purposes, we simply adopt the notion of latent quality as a modeling assumption. In the next section, we discuss the possibility of doing away with the notion of quality altogether.

Note that even without an argument that selecting for quality is an empirically reasonable description of peer review, it seems clear that scientists and editors themselves view it this way and discuss it in these terms. From this perspective, our models simply hold up a mirror, pointing out in abstract terms and under minimal auxiliary assumptions that the following natural idea is false: If editors select for quality in an unbiased way, then they will treat different research programs equally. This reveals something important, we think, about the baseline case of selecting for quality, independently of whether our models describe phenomena that occur in practice.

If we accept the notion of quality in some form, it is still not clear what we should take to be expressed by the quality scale. The model assumes that the conditional probability of acceptance $\Pr(A \mid q)$ increases monotonically in q, but other than that, it is a matter for further discussion how to interpret it. The quality of a paper might be the long-run importance of the paper in a discipline, as a social fact, or perhaps the contribution that a paper makes to the development of a discipline towards some goal. Referees and editors will rank papers according to several criteria, which are then compressed into a binary judgment. It is to some extent an empirical question what weighted combination of criteria is best taken as the perceived quality.[7]

For our concerns, a particularly salient consideration is that editors and referees might take the novelty or originality of the paper as one of the criteria. That is, a paper may receive a high quality ranking because it brings a fresh perspective to a discipline, e.g., by working from within a new research program. As we have argued, at the level of a scientific discipline, epistemic diversity is a stand-alone virtue because it improves the versatility and hence resilience of the discipline as a whole. However, if novelty by itself enhances the quality of individual papers, this would presumably undercut our main conclusions about the differences between established and novel programs.

As indicated before, our results present a baseline case in which editors do not factor in such global considerations when judging individual papers for their journal. Their primary goal is to maintain their journal's status, and therefore to publish papers that offer good descriptions, reliable predictions, and convincing explanations. In our model, novelty of perspective may still contribute to the quality of a paper in a derivative sense, in that it may occasion benefits for the individual paper that matter to an editor, e.g., when the novel perspective makes for better descriptive, predictive, or explanatory properties. Whether it will feature as a global consideration in its own right, is once again an empirical issue. It is not taken into consideration in the model we have presented.

We turn to the specific modeling assumptions that drive our results. For the first result, we assumed that the editor will be more acquainted with authors, networks, or research groups from an established program than from a new one. This seems to be fairly straightforward:

6. Our reasons to think this relate to a phenomenon known from psychometrics, the "positive manifold" (Spearman 1904). Insofar as the various quality aspects of a paper will correlate positively — and we think they will — we can typically include a single latent variable as a modeling tool and interpret it as paper quality, without committing to its existence or causal efficacy (cf. van der Maas et al. 2006).

7. The familiar paradoxes of voting theory loom here: It may be impossible to aggregate scores on criteria in a way that avoids a "dictatorship" of one quality criterion.

A more established program will have had more exposure and more time, and it is also more likely that the editor herself is associated with it. For the second result, however, the key assumption is that the average quality of papers from the more established program is higher. This assumption is far less straightforward, but we believe it can be motivated.

Characteristics that determine the quality of a paper are its descriptive, predictive, and explanatory characteristics. They are in turn determined by the so-called positive heuristics of the program from which a paper originates, i.e., its core assumptions, and furthermore by the skill sets and the institutional and financial support of the researchers. These latter characteristics underpin the differences in the average quality of the programs: More established programs will have more social and monetary capital to make their research a success, they will have more developed methods and techniques, and also better training facilities. Additionally, those programs will be better equipped to recognize, support, and signal quality and talent. If the core assumptions of the novel program are superior, then we might hope that this will eventually come to light. But the novel program starts at a disadvantage.

5. Counteracting Statistical Biases

In the foregoing, we offered a critical assessment of our models, and argued that statistical biases might well be a reality. We readily admit that the models are not a complete description of editorial decision making. The statistical biases that we identify will be mixed in with biases and mechanisms that we have not described. However, this does not take away from the need to counteract the biases identified. As long as we believe that the models capture certain aspects of real editorial practice, the statistical biases might indeed obtain, even if they obtain alongside others. Hence, we have reason to look for ways to respond. We devote the remainder of our paper to the question of how we might do this.

Before we consider our options, it deserves emphasis that our results are still valuable if further empirical investigation reveals that the modeling assumptions are too idealized, and that they therefore never obtain. That is, our results are also informative when they are not merely incomplete, as already discussed, but empirically false. In that case, they still present a principled argument against the feasibility of a particular ideal of editorial decision making. This makes them informative for our editorial practices in a derivative sense. They tell us that, as a baseline case, a strict focus on individual paper quality may be detrimental to program diversity, and that differential treatment at some level seems inevitable.

Assuming now that we have to counteract these biases, what can we do? One response to the first of the two results was already discussed at the start of section 3. This bias can be prevented by disallowing the information asymmetry required for the result. We could demand that no prior information about the author of a paper may be taken into account in evaluating it, analogously to the standing practice in criminal prosecution and psychological testing for the purpose of selection. One way to achieve this is by employing triple-anonymous review, but it should be noted that the editor then foregoes information that would help her improve the average quality of papers in her journal. Another option is that we remedy the information asymmetry between programs by working with an editorial team that reflects the mix of research programs in the discipline.

Owing to our second result, however, this approach fails to rule out all threats to epistemic diversity. We readily admit that the assumption of a lower average quality for the new program will not always hold, but we believe we have motivated it sufficiently to say it holds sometimes, so that the statistical bias in the editorial process can occur. We also believe that it would be a mistake to consider this kind of bias relatively harmless, or even reasonable in the light of the latent differences between the programs. It is to be expected that the program producing lower quality work gets this work published less easily, but

it is far more worrisome that science's selection mechanisms work less well for that program.[8]

One more-or-less direct repair can be constructed. The root cause of the differences in error rates is that the novel program has proportionally more papers that are near the quality cut-off point for inclusion in the journal. Accordingly, we can counteract the bias by directing more reviewer efforts towards papers that are borderline cases. To some extent, this is already the standing practice, or so we think. A problematic consequence of this is perhaps that this creates an asymmetry between the two programs of a different nature: The editorial office will spend more of its reviewing resources on the novel program (cf. the analogous discussion in Borsboom et al. 2008). This will be acceptable to some, but others may feel that the persistence of biases invites us to search farther afield. In particular, we might hope to eliminate the implicit adoption of a notion of quality in editorial decision making.

What we are taking into consideration here is a more far-reaching re-evaluation of the system of science. Scientific publication is a process of regulated information sharing. Depending on what goals we take this information sharing to have, it may well turn out that it is better served by a system like ArXiv than by centralized collection and curation. To find out about this, we need to confront our models with empirical fact and evaluate the merits and defects of the various formats for information sharing in science (in a forthcoming paper, Heesen and Bright attempt to do this). Indeed, the ultimate resolution of threats to epistemic diversity through biases in editorial decision making might turn out to be a truly radical one: to do away with editor decisions altogether. Depending on the details of such a system, new problems will undoubtedly emerge, but we may hope that statistical Matthew effects are not among them. Even among the authors of this article, the debate over the merits and defects of thoroughly overhauling editorial processes continues.[9]

8. That such biases are to be taken seriously is underscored by the public debate around fairness in AI and the scientific work that it has promoted. Kleinberg et al. (2017), for instance, prove a version of our theorem 3, and suggest that it is the main driver behind the unfairness of a system that estimates recidivism risks for the US criminal courts (Angwin et al. 2016).

Appendix A. Proof of the Value of Knowing the Author

Our proof relies on the following well-known result.

Theorem 6 (Good (1967))**.** *Given some choice problem, let D be a decision that maximizes expected utility relative to some prior beliefs and a utility function u. Let K be a random variable and let $D(K)$ be a decision that maximizes expected utility relative to the posterior beliefs (obtained from the prior beliefs by conditioning on the outcome of K) and utility u. Then*

$$\mathbb{E}_K[\mathbb{E}[u(D(K))]] \geq \mathbb{E}[u(D)].$$

Moreover, the foregoing inequality is strict if there is a set of outcomes K_0 such that $\Pr(K \in K_0) > 0$ and $\mathbb{E}[u(D(k))] > \mathbb{E}[u(D)]$ for all $k \in K_0$ (i.e., decision D no longer maximizes expected utility if outcome k is observed).

In our model, we make a distinction between scientists known to the editor and scientists unknown to the editor, where knowing a scientist is represented as knowing the value of some random variable K that is potentially relevant to evaluating the quality of the scientist's paper. Let $D(K, R)$ be the decision taken by the editor if she knows the scientist's characteristics K and the reviewer report R and let $D(R)$ be the decision if the scientist is unknown, i.e., only the reviewer report R is known. Applying Good's theorem to our model yields the following.

9. We thank Cailin O'Connor, Liam Bright, Mike Schneider, Leah Henderson, Hannah Rubin, Herbert Hoijtink, Philip Kitcher, as well as audiences in Bristol, Hannover, Bochum, Rome, Cologne, and Seattle for valuable comments and discussion. RH's research was supported by the Netherlands Organisation for Scientific Research (NWO) under grant 016.Veni.195.141 and by the Leverhulme Trust and the Isaac Newton Trust under an Early Career Fellowship. To contact the authors, please write to remco.heesen@uwa.edu.au or j.w.romeijn@rug.nl.

Theorem 7. *Assume that there exists a set S of joint outcomes for K and R (i.e., members of S are pairs (k, r) where k is a possible outcome of K and r is a possible outcome of R) such that $D(k, r) \neq D(r)$ for all $(k, r) \in S$ and $\Pr((K, R) \in S) > 0$. Then*

$$\begin{aligned} \Pr(D(K,R) = A) > \Pr(D(R) = A) \quad \textit{or} \\ \mathbb{E}[q \mid D(K,R) = A] > \mathbb{E}[q \mid D(R) = A]. \end{aligned}$$

Proof. From theorem 6, we get that

$$\mathbb{E}_K[\mathbb{E}[u(D(K,R))]] \geq \mathbb{E}[u(D(R))],$$

with strict inequality if there is a set of outcomes for K with positive measure such that $\mathbb{E}[u(D(k,R))] > \mathbb{E}[u(D(R))]$ for all k in that set. The theorem assumes that such sets of outcomes exist, so we have strict inequality in the above.

From the definition of u, we know that $u(D(R)) = q$ if $D(R) = A$ and $u(D(R)) = q^*$ otherwise. Hence

$$\begin{aligned} \mathbb{E}[u(D(R))] &= \mathbb{E}[q \mid D(R) = A] \Pr(D(R) = A) + q^* \Pr(D(R) = \neg A) \\ &= q^* + \mathbb{E}[q - q^* \mid D(R) = A] \Pr(D(R) = A). \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}_K[\mathbb{E}[u(D(K,R))]] &= \mathbb{E}_K[\mathbb{E}[q \mid D(K,R) = A]] \Pr(D(K,R) = A) \\ &\quad + q^* \Pr(D(K,R) = \neg A) \\ &= q^* + \mathbb{E}_K[\mathbb{E}[q - q^* \mid D(K,R) = A]] \Pr(D(K,R) = A). \end{aligned}$$

The inequality obtained from theorem 6 entails

$$\begin{aligned} \mathbb{E}_K[\mathbb{E}[q - q^* \mid D(K,R) = A]] > \mathbb{E}[q - q^* \mid D(R) = A] \quad \text{or} \\ \Pr(D(K,R) = A) > \Pr(D(R) = A). \end{aligned}$$

Since q^* is a constant, the former inequality is equivalent to the one stated in the theorem. □

The above theorem assumes that there exists a set of outcomes S for K and R of positive probability such that $D(k, r) \neq D(r)$ for all $(k, r) \in S$. This is a more formally precise statement of the assumption made in theorem 2 that knowing the characteristics of a scientist sometimes makes a difference to the editor's decision. Theorem 2 follows as a corollary of theorem 7.

Proof of theorem 2. Conditional on whether or not the editor knows the characteristics of the scientist who wrote the paper, knowing which research program the paper belongs to is completely irrelevant: Both the quality distribution and the decision procedure used are identical for research programs H and L. It follows that both the probability of acceptance and the average quality of published papers are the same, i.e.,

$$\begin{aligned} \Pr(D(K,R) = A \mid H) &= \Pr(D(K,R) = A \mid L), \\ \Pr(D(R) = A \mid H) &= \Pr(D(R) = A \mid L), \\ \mathbb{E}[q \mid D(K,R) = A, H] &= \mathbb{E}[q \mid D(K,R) = A, L], \\ \mathbb{E}[q \mid D(R) = A, H] &= \mathbb{E}[q \mid D(R) = A, L]. \end{aligned}$$

From theorem 7, we get that either the first of the above four lines is greater than the second, or the third line is greater than the fourth. Let p_{KH} denote the proportion of scientists in research program H known to the editor and let p_{KL} denote the proportion of scientists in research program L known to the editor. Then

$$\begin{aligned} \Pr(A \mid H) &= p_{KH} \Pr(D(K,R) = A \mid H) \\ &\quad + (1 - p_{KH}) \Pr(D(R) = A \mid H), \\ \mathbb{E}[q \mid A, H] &= p_{KH} \mathbb{E}[q \mid D(K,R) = A, H] \\ &\quad + (1 - p_{KH}) \mathbb{E}[q \mid D(R) = A, H], \end{aligned}$$

and similarly for $\Pr(A \mid L)$ and $\mathbb{E}[q \mid A, L]$. The result follows from the assumption that $p_{KH} > p_{KL}$. □

Appendix B. Proof of the Consequences of Latent Differences

For our purposes, the following characterization of log-concave densities is key. See Saumard and Wellner (2014, p. 97) for a proof.

Theorem 8. *Density function f is log-concave if and only if the family of densities f_G defined by $f_G(q) := f(q - \mu_G)$ has monotone likelihood ratios, i.e.,*

$$\frac{f_H(q)}{f_L(q)} = \frac{f(q-\mu_H)}{f(q-\mu_L)} \geq \frac{f(q'-\mu_H)}{f(q'-\mu_L)} = \frac{f_H(q')}{f_L(q')}$$

whenever $q > q'$, $\mu_H > \mu_L$, $f_L(q) > 0$, and $f_L(q') > 0$.

The above theorem is used in the proof of our main result.

Proof of theorem 3. Let $f_H = F_H'$ and $f_L = F_L'$ be the density functions for the latent in the two groups. We first consider the distribution of quality conditional upon acceptance. Note that

$$\Pr(q > t \mid A, H) = \frac{\Pr(q > t, A \mid H)}{\Pr(A \mid H)} = \frac{\int_t^\infty a(q) f_H(q)\, dq}{\int_{-\infty}^\infty a(q) f_H(q)\, dq},$$

$$\Pr(q > t \mid A, L) = \frac{\int_t^\infty a(q) f_L(q)\, dq}{\int_{-\infty}^\infty a(q) f_L(q)\, dq}.$$

Consider the following special cases:

- If $c + \mu_L \leq t < c + \mu_H$, we are done immediately because $\Pr(q > t \mid A, H) > 0 = \Pr(q > t \mid A, L)$.
- If $b + \mu_L < t \leq b + \mu_H$, we are done immediately because $\Pr(q > t \mid A, H) = 1 > \Pr(q > t \mid A, L)$.
- If $c + \mu_L \leq b + \mu_H$, we are done immediately because for any value of t either $\Pr(q > t \mid A, H) = 1$ or $\Pr(q > t \mid A, L) = 0$.

So for the remainder of the proof, we need only consider the case where $b + \mu_H < t < c + \mu_L$. By theorem 8, f_H / f_L is a non-decreasing function of q whenever it exists. This function exists for all q such that $f_L(q) > 0$, so in particular for $q \in (t, c + \mu_L)$. Thus

$$\begin{aligned}\Pr(q > t \mid A, H) &= \frac{\int_t^{c+\mu_L} a(q) f_H(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) f_H(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq} \\ &= \frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}.\end{aligned}$$

Since $b + \mu_H < t$, the numerator of this fraction is strictly smaller than the denominator, i.e., $\Pr(q > t \mid A, H) < 1$. It follows that

$$\Pr(q > t \mid A, H) \geq \frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq},$$

with strict inequality if $c < \infty$. Hence it suffices to show that

$$\frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_t^{c+\mu_L} a(q) f_L(q)\, dq} \geq \frac{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) f_L(q)\, dq},$$

with strict inequality if $c = \infty$. Let X be a random variable whose density function is given by

$$f_X(x) = \frac{a(x)f_L(x)}{\int_{b+\mu_L}^{c+\mu_L} a(u)f_L(u)\,du}$$

for all x. Then the above inequality is equivalent to

$$\mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X > t\right] \geq \mathbb{E}\left[\frac{f_H(X)}{f_L(X)}\right].$$

This inequality holds because f_H/f_L is non-decreasing by theorem 8. It remains to show that this inequality holds strictly if $c = \infty$. Equivalently, it remains to show that, if $c = \infty$,

$$\mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X > t\right] > \mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X \leq t\right].$$

Because f_H/f_L is non-decreasing, $f_H(t)/f_L(t)$ is a lower bound for the left-hand side of this inequality, and an upper bound for the right-hand side. Since t is assumed to be in the support of both f_H and f_L, $f_H(t)/f_L(t) > 0$.

If $b > -\infty$, then for $b + \mu_L < x < b + \mu_H$, we have $f_H(x) = 0$. Hence, conditional on $X < t$, $f_H(X)/f_L(X) = 0$ with positive probability, and thus the expectation on the right-hand side must be strictly smaller than $f_H(t)/f_L(t)$.

If $b = -\infty$, then the inequality is strict unless $f_H(x)/f_L(x) = f_H(t)/f_L(t)$ for all $x \in \mathbb{R}$. But that happens only if $f_H = f_L$, i.e., if $F_H = F_L$. But we know that $F_H \neq F_L$ because these distributions are obtained from F by adding different constants $\mu_H \neq \mu_L$.

This concludes the proof for the distribution of quality given acceptance. Now consider the probability of acceptance given $q > t$.

$$\Pr(A \mid q > t, H) = \frac{\Pr(q > t, A \mid H)}{\Pr(q > t \mid H)} = \frac{\int_t^\infty a(q)f_H(q)\,dq}{\int_t^\infty f_H(q)\,dq}$$

$$= \mathbb{E}[a(q) \mid q > t, H],$$

$$\Pr(A \mid q > t, L) = \frac{\int_t^\infty a(q)f_L(q)\,dq}{\int_t^\infty f_L(q)\,dq}.$$

Note that if $c + \mu_L \leq t < c + \mu_H$, then $\Pr(q > t \mid L) = 0$. This would mean that $\Pr(A \mid q > t, L)$ is not defined, so we set this case aside and suppose that $t < c + \mu_L$.

We may write $\mathbb{E}[a(q) \mid q > t, H]$ as a weighted average of $\mathbb{E}[a(q) \mid q > c + \mu_L, H]$ and $\mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H]$. Since a is an increasing function,

$$\mathbb{E}[a(q) \mid q > c + \mu_L, H] > a(c + \mu_L) > \mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H].$$

Hence

$$\Pr(A \mid q > t, H) \geq \mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H]$$

$$= \frac{\int_t^{c+\mu_L} a(q)f_H(q)\,dq}{\int_t^{c+\mu_L} f_H(q)\,dq}$$

$$= \frac{\int_t^{c+\mu_L} a(q)\frac{f_H(q)}{f_L(q)}f_L(q)\,dq}{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)}f_L(q)\,dq},$$

with strict inequality if $c < \infty$. Then it suffices to show that

$$\begin{aligned}\mathbb{E}\left[\frac{f_H(Y)}{f_L(Y)}\right] &= \frac{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)} a(q) f_L(q)\, dq}{\int_t^{c+\mu_L} a(q) f_L(q)\, dq} \\ &\geq \frac{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_t^{c+\mu_L} f_L(q)\, dq} \\ &= \mathbb{E}\left[\frac{f_H(Z)}{f_L(Z)}\right],\end{aligned}$$

where Y and Z's density functions are given respectively by

$$f_Y(x) = \begin{cases} \dfrac{a(x) f_L(x)}{\int_t^{c+\mu_L} a(u) f_L(u)\, du} & \text{if } x > t, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_Z(x) = \begin{cases} \dfrac{f_L(x)}{\int_t^{c+\mu_L} f_L(u)\, du} & \text{if } x > t, \\ 0 & \text{otherwise.} \end{cases}$$

Note that whenever $x > t$,

$$\frac{f_Y(x)}{f_Z(x)} \propto a(x),$$

which is increasing in x by assumption. So Y has relatively higher density for high values. Since, moreover, f_H/f_L is non-decreasing, it follows that

$$\mathbb{E}\left[\frac{f_H(Y)}{f_L(Y)}\right] \geq \mathbb{E}\left[\frac{f_H(Z)}{f_L(Z)}\right].$$

The inequality is an equality only if $f_H(q)/f_L(q) = f_H(t)/f_L(t)$ for all $q > t$. This happens if and only if $f(q) \propto \exp\{-q\}$. □

Proof of corollary 4. By theorem 3, we have

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L)$$

for all $t \in (b + \mu_L, c + \mu_H)$. For any t outside this interval, the above inequality is an equality (both probabilities are one if $t \leq b + \mu_L$ and zero otherwise). Thus the distribution $q \mid A, H$ first-order stochastically dominates the distribution of $q \mid A, L$. It follows that

$$\mathbb{E}\left[q \mid A, H\right] > \mathbb{E}\left[q \mid A, L\right].$$

We could use the other inequality from theorem 3 to establish the inequality in acceptance rates, but then we would need to worry about the special case where the right tail of f is exponential. Instead, we provide a simple direct proof of the inequality in acceptance rates, which also shows that the assumption that f is log-concave is superfluous.

$$\begin{aligned}\Pr(A \mid H) &= \int_{b+\mu_H}^{c+\mu_H} a(q) f(q - \mu_H)\, dq \\ &= \int_{b+\mu_L}^{c+\mu_L} a(u + \mu_H - \mu_L) f(u - \mu_L)\, du \\ &> \int_{b+\mu_L}^{c+\mu_L} a(u) f(u - \mu_L)\, du = \Pr(A \mid L).\end{aligned}$$

□

Proof of theorem 5. By the chain rule, F_H and F_L are differentiable and their densities are given by

$$f_H(q) = \frac{1}{\sigma_H} f\left(\frac{q-\mu_H}{\sigma_H}\right) \quad \text{and} \quad f_L(q) = \frac{1}{\sigma_L} f\left(\frac{q-\mu_L}{\sigma_L}\right)$$

for all q.

Consider the probability that a paper from research program H is accepted and its quality q exceeds t. Using the substitution $q = t + \frac{\sigma_H}{\sigma_L}(u-t)$ we find:

$$\begin{aligned}
\Pr(q > t, A \mid H) &= \int_t^{\sigma_H c + \mu_H} a(q) \frac{1}{\sigma_H} f\left(\frac{q-\mu_H}{\sigma_H}\right) dq \\
&= \int_t^{\sigma_L c + t + \frac{\sigma_L}{\sigma_H}(\mu_H - t)} a\left(t + \frac{\sigma_H}{\sigma_L}(u-t)\right) \\
&\qquad \cdot \frac{1}{\sigma_L} f\left(\frac{1}{\sigma_L}\left(u - t - \frac{\sigma_L}{\sigma_H}(\mu_H - t)\right)\right) du \\
&= \int_t^{\sigma_L c + \mu'} a\left(t + \frac{\sigma_H}{\sigma_L}(u-t)\right) g(u-\mu')\, du,
\end{aligned}$$

where g is the function given by $g(x) = f(x/\sigma_L)/\sigma_L$ and $\mu' = t + \frac{\sigma_L}{\sigma_H}(\mu_H - t)$. Since $u > t$ and $\sigma_H > \sigma_L$, we have $t + \frac{\sigma_H}{\sigma_L}(u-t) > u$. Using the fact that a is an increasing function:

$$\Pr(q > t, A \mid H) > \int_t^{\sigma_L c + \mu'} a(u) g(u-\mu')\, du.$$

Analogously, we find that

$$\Pr(q \leq t, A \mid H) < \int_{\sigma_L b + \mu'}^{t} a(u) g(u-\mu')\, du.$$

Applying these two inequalities yields

$$\begin{aligned}
\Pr(q > t \mid A, H) &= \frac{\Pr(q > t, A \mid H)}{\Pr(q > t, A \mid H) + \Pr(q \leq t, A \mid H)} \\
&> \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u-\mu')\, du}{\int_{\sigma_L b + \mu'}^{\sigma_L c + \mu'} a(u) g(u-\mu')\, du}.
\end{aligned}$$

Note that the function g is itself a density function: In particular, if f is the density function of some random variable X, then g is the density function of the random variable $\sigma_L X$. Since f is log-concave, and log-concavity is preserved by affine transformations (Saumard and Wellner 2014, p. 57), g is also log-concave.

But then we can apply theorem 3! In particular, the condition $(\mu_H - t)/\sigma_H \geq (\mu_L - t)/\sigma_L$ is equivalent to $\mu' \geq \mu_L$. Hence by theorem 3:

$$\begin{aligned}
\Pr(q > t \mid A, H) &> \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u-\mu')\, du}{\int_{\sigma_L b + \mu'}^{\sigma_L c + \mu'} a(u) g(u-\mu')\, du} \\
&\geq \frac{\int_t^{\sigma_L c + \mu_L} a(u) g(u-\mu_L)\, du}{\int_{\sigma_L b + \mu_L}^{\sigma_L c + \mu_L} a(u) g(u-\mu_L)\, du} \\
&= \frac{\int_t^{\sigma_L c + \mu_L} a(u) \frac{1}{\sigma_L} f\left(\frac{u-\mu_L}{\sigma_L}\right) du}{\int_{\sigma_L b + \mu_L}^{\sigma_L c + \mu_L} a(u) \frac{1}{\sigma_L} f\left(\frac{u-\mu_L}{\sigma_L}\right) du} = \Pr(q > t \mid A, L).
\end{aligned}$$

This proves the first inequality. The second inequality is quite similar. Consider the probability that the quality of a paper from research program H exceeds t. Using again the substitution $q = t + \frac{\sigma_H}{\sigma_L}(u-t)$, we

find:

$$\Pr(q > t \mid H) = \int_t^{\sigma_H c + \mu_H} \frac{1}{\sigma_H} f\left(\frac{q - \mu_H}{\sigma_H}\right) dq$$

$$= \int_t^{\sigma_L c + \mu'} g(u - \mu')\, du.$$

Combining this with the result for $\Pr(q > t, A \mid H)$ from the first half of the proof, we see that

$$\Pr(A \mid q > t, H) = \frac{\Pr(q > t, A \mid H)}{\Pr(q > t \mid H)} > \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u - \mu')\, du}{\int_t^{\sigma_L c + \mu'} g(u - \mu')\, du}.$$

Since g is log-concave and $\mu' \geq \mu_L$, we can apply theorem 3 to get

$$\Pr(A \mid q > t, H) > \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u - \mu')\, du}{\int_t^{\sigma_L c + \mu'} g(u - \mu')\, du}$$

$$\geq \frac{\int_t^{\sigma_L c + \mu_L} a(u) g(u - \mu_L)\, du}{\int_t^{\sigma_L c + \mu_L} g(u - \mu_L)\, du}$$

$$= \frac{\int_t^{\sigma_L c + \mu_L} a(u) \frac{1}{\sigma_L} f\left(\frac{u - \mu_L}{\sigma_L}\right) du}{\int_t^{\sigma_L c + \mu_L} \frac{1}{\sigma_L} f\left(\frac{u - \mu_L}{\sigma_L}\right) du} = \Pr(A \mid q > t, L). \square$$

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Propublica*, May 23, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed July 4, 2018.

Jessica L. Blackburn and Milton D. Hakel. An examination of sources of peer-review bias. *Psychological Science*, 17(5):378–382, 2006. URL http://dx.doi.org/10.1111/j.1467-9280.2006.01715.x.

Denny Borsboom, Jan-Willem Romeijn, and Jelte M. Wicherts. Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, 13(2):75–98, 2008. URL http://dx.doi.org/10.1037/1082-989X.13.2.75.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. URL http://dx.doi.org/10.1023/A:1016082229411.

E. Ernst, K. L. Resch, and E. M. Uher. Reviewer bias. *Annals of Internal Medicine*, 116(11):958, 1992. URL http://dx.doi.org/10.7326/0003-4819-116-11-958_2.

Paul Feyerabend. *Against Method*. New Left Books, London, 1975.

I. J. Good. On the principle of total evidence. *The British Journal for the Philosophy of Science*, 17(4):319–321, 1967. URL http://www.jstor.org/stable/686773.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen and Liam Kofi Bright. Is peer review a good idea? *The British Journal for the Philosophy of Science*, forthcoming. URL http://dx.doi.org/10.1093/bjps/axz029.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004. URL http://dx.doi.org/10.1073/pnas.0403723101.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. URL http://www.jstor.org/stable/2026796.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, pages 43:1–43:23, 2017. URL https://arxiv.org/abs/1609.05807.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994. URL http://www.jstor.org/stable/2138799.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, Cambridge, 1978.

Carole J. Lee and Christian D. Schunn. Philosophy journal practices and opportunities for bias. *American Philosophical Association Newsletter on Feminism and Philosophy*, 10(1):5–10, 2010. URL http://cdn.ymaws.com/www.apaonline.org/resource/collection/D03EBDAB-82D7-4B28-B897-C050FDC1ACB4/v10n1Feminism.pdf.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. URL http://dx.doi.org/10.1002/asi.22784.

Helen E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, 1990.

Terttu Luukkonen. Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation*, 21(1):48–60, 2012. URL http://dx.doi.org/10.1093/reseval/rvs001.

Michael J. Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1(2):161–175, 1977. URL http://dx.doi.org/10.1007/BF01173636.

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968. URL http://www.jstor.org/stable/1723414.

K. I. Resch, E. Ernst, and J. Garrow. A randomized controlled study of reviewer bias against an unconventional therapy. *Journal of the Royal Society of Medicine*, 93(4):164–167, 2000. URL http://dx.doi.org/10.1177/014107680009300402.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014. URL http://dx.doi.org/10.1214/14-SS107.

Miriam Solomon. *Making Medical Knowledge*. Oxford University Press, Oxford, 2015.

C. Spearman. "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. URL http://dx.doi.org/10.2307/1412107.

Shelley L. Tremain. *Foucault and Feminist Philosophy of Disability*. University of Michigan Press, Ann Arbor, 2017.

Han L. J. van der Maas, Conor V. Dolan, Raoul P. P. P. Grasman, Jelte M. Wicherts, Hilde M. Huizenga, and Maartje E. J. Raijmakers. A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4):842–861, 2006. URL http://dx.doi.org/10.1037/0033-295X.113.4.842.

Alison Wylie. Community-based collaborative archaeology. In Nancy Cartwright and Eleonora Montuschi, editors, *Philosophy of Social Science: A New Introduction*, chapter 4, pages 68–82. Oxford University Press, Oxford, 2014.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010. URL http://dx.doi.org/10.1007/s10670-009-9194-6.